Accuracy of Speed Zone Recorded in the Victorian Police-Reported Crash Database Differs by Speed Zone in Metropolitan Areas

Karen Stephan, Stuart Newstead

Monash University Accident Research Centre

Abstract

We investigated the accuracy (overall, sensitivity and positive predictive value (PPV)) of the speed zone recorded for the location of crashes in a sample of Victorian police-reported crash data. Data on crashes that occurred on 111 road segments (where the true speed limit was known) in the Melbourne metropolitan area were obtained. The speed zone recorded for each crash location was compared to the true speed limit. Speed zone was not recorded correctly in almost one-quarter of crashes. Sensitivity varied by speed zone and was highest for crashes that occurred in 60 km/h zones. PPV was extremely low for crashes recorded as occurring in 50 km/h zones; almost 90% did not occur in 50 km/h zones. Misclassification of speed zone was less frequent when police attended the scene. The misclassification affected estimates of association between speed zone and crash severity. Implications and recommendations for rectifying the problem are discussed.

Background

It is essential to have accurate and reliable data on traffic crashes to measure road safety, including changes over time. Administrative crash databases held by road authorities are an important source of data that are used widely by a range of organizations; from Government departments and the police to academic researchers and the general public. Such data are used to measure the size of the road safety problem, to identify risk factors, to evaluate countermeasures and to monitor performance with the goal of informing evidence-based policy and strategy to reduce crashes and injuries (World Health Organization, 2010). It is sometimes difficult, however, to establish the accuracy of the data contained in administrative crash databases.

Much of the data on crashes is sourced from reports completed by police. For example, in Victoria, Vicroads administer a database of police-reported crashes (Vicroads, 2008) that includes crashes which resulted in the death of any person within 30 days of the crash or where the police identified one of the road users as injured. The database includes information on the crash circumstances, crash location, road users involved, vehicles involved, the road environment and road and weather conditions. The majority of the data in the database are recorded by police who attend the crash scene. In cases where a person attends a police station to report a crash, the police officer will collect information about the crash from the person reporting the crash. Reliance on police and/or the person reporting the crash to recall details of the crash (including location, road users, vehicles, road environment and conditions) means that the data are subject to error. There is potential in future, however, to improve data quality and relieve police of the requirement to collect data that are available from other sources, for example, geo-spatial databases of road infrastructure and speed zones.

Some examples of how the Victorian crash data have been used for road safety purposes include describing the nature of the crashes, both for the whole population and for particular subgroups of road users, for example young drivers and cyclists (e.g. Garratt, Johnson & Cubis, 2015; Buckis, Lenne, Stephens, Bingham & Fitzharris, 2016; Vicroads, 2016), identifying crash risk factors relating to the road users, the vehicles, the road and the environment (e.g. Boufous, de Rome, Senserrick & Ivers, 2012) and in ongoing analyses and monitoring of vehicle crashworthiness for the Used Car Safety Ratings (e.g. Newstead, Watson, L.M. & Cameron, 2013). Characteristics of the crash location, including speed zone, are frequently considered in such studies; from describing the proportion of crashes that occur in each speed zone, to measuring the relationship between speed zone and crash severity. Results of these studies have been used to identify speed zones that are riskier for different types of road users, and to recommend treatments and countermeasures in different speed zones to address the risk. Yet it is unclear how accurately the speed zone for crash locations is recorded in the Victorian crash data.

Errors in the crash data, in terms of misclassification of crashes, are a problem because they can influence the validity of the research results. The effect will depend on whether the misclassification is non-differential or differential. For example, consider a study investigating the association between speed zone and crash severity. If the probability that the speed zone recorded for a crash location is misclassified does not differ across accident severities (e.g. 10% of crashes of each severity are recorded incorrectly as occurring in another speed zone: non-differential misclassification), then the estimate of the association between speed zone and crash severity using the non-differentially misclassified data will be closer to the null (that is, no effect) than if the speed zones were not misclassified. That is, non-differential misclassification always dilutes the estimates of the association between an exposure and an outcome (Hennekens & Buring, 1987) and leads to a more conservative estimate. Differential misclassification, however, is a bigger problem for the validity of the results of research. If the probability that the speed zone recorded for a crash location is misclassified differs according to crash severity (for example, if misclassification of speed zone is more likely for other injury crashes than more severe crashes), then the estimate of the association between speed zone and crash severity using the differentially misclassified data will be incorrect, but, the estimate of association may be larger or smaller than the true estimate. That is, differential misclassification affects the estimate of association, but the direction of the bias cannot be predicted (Hennekens & Buring, 1987), and therefore, a particular speed zone may look riskier, or safer, than is truly the case. Misclassification of speed zone may be more likely for crashes that occur in urban areas than in rural areas, due to the greater number of roads of different speed zones within small areas. It is therefore important to investigate whether misclassification is present in crash data, particularly for urban crashes.

To measure the accuracy of the data held in administrative crash databases, we must compare the data in the crash database to data from a source known to be accurate. This is difficult for many of the variables in the crash data. It is possible, however, for speed zone, because the speed zones of all roads in Victoria have been mapped. We can therefore compare the speed recorded for each crash location to the true speed zone at that crash location.

There are several measures of accuracy that are useful for measuring the accuracy of the speed zone information in the Victorian crash data. The first is the proportion of crashes for which the speed is correctly recorded (overall measure of accuracy) and incorrectly recorded (or misclassified). It is also of interest to determine if the accuracy differs by speed zone; that is, the proportion of crashes that occur in each speed zone that are recorded as occurring in that speed zone (sensitivity). Finally, for those researchers and practitioners who use the crash data, it is important to know how much they can trust the data in the database. For this, we can measure the positive predictive value (PPV); the proportion of crashes recorded as occurring in a particular speed zone that actually occurred in that speed zone.

The aim of this study was to investigate how accurately the speed zone in which crashes occurred was recorded in Victorian police-reported crash data for urban roads. Specifically, we aimed to measure the overall accuracy, the sensitivity and the PPV of the reported speed zone. Secondary aims were to identify factors related to the accuracy of speed zone recording and to determine what effect the misclassification of speed zone had on the relationship between speed zone and accident severity.

Method

Road segments

As part of a larger project investigating risk factors for crashes in complex urban areas (Stephan, 2015), 142 strip shopping arterial road segments in metropolitan Melbourne of at least 200 metres in length, composed predominantly of retail buildings on one or both sides of the road, were selected. For the purposes of the current investigation, road segments with variable speed limits (either for schools (n=17) or for strip shopping (n=14)) were excluded. Therefore, 111 road segments were included in this study.

Data sources

Data on the crashes that occurred on the 111 road segments between 2005 and 2009 were obtained from the Vicroads database of police-reported casualty crashes (Vicroads, 2008). This includes crashes which resulted in the death of any person within 30 days of the crash or where the police identified one of the road users as injured. The database includes information on the crash circumstances, location, road users involved, vehicles involved and the environment. These data are recorded by police. The speed zone of the road on which the crash occurred is included among the data collected by police for each crash, either through attending the crash scene or from information obtained from the person reporting the crash at a police station.

Data on the true speed zone of each road segment were obtained from Vicroads. Vicroads staff referred to historical records to confirm there were no speed limit changes on the road segments between 2005 and 2009.

Data analysis

Crashes that occurred on midblock road segments (not at an intersection) were identified. Intersection crashes were excluded because we could not be sure if the speed zone was recorded for the road segment of interest, or the intersecting road (for which we did not have information on the true speed zone). For each road segment, the number of midblock crashes that were recorded in each speed zone was counted and compared to the true speed zone for that road segment. Data were collated across speed zones and the sensitivity (% of crashes in each speed zone that were recorded correctly) and the PPV (% of crashes recorded as being in a particular speed zone, that were actually in that speed zone) were calculated for each speed zone.

Each crash was classified according to whether or not the speed zone was recorded correctly and further investigation was carried out to determine the factors associated with whether or not the speed zone at the crash site was recorded correctly. Separate simple logistic regression models were used to estimate the relationship between whether the speed zone was recorded correctly (binary outcome) and the following predictors:

- The true speed zone (that is, whether sensitivity varied significantly across speed zones)
- The recorded speed zone (that is, whether the PPV of the recorded speed limit varied significantly across speed zones)
- Accident severity (as defined in the police report); and
- Whether or not the police attended the crash scene

A multiple logistic regression model was used to estimate the independent contributions of police attendance and accident severity to the accuracy of the recorded speed zone.

Finally, in order to demonstrate the effect of inaccuracies in recording of speed zone, the relationship between the speed zone at the crash location and the accident severity (binary outcome: whether or not a road user was killed or seriously injured in the crash, or received less severe injuries) was estimated using simple logistic regression. Two models were developed: the first using the true speed zone and the second using the speed zone as recorded in the police-reported crash data (some of which were inaccurate).

Results

Accuracy of recorded speed zone

There were 655 crashes that occurred at midblock locations on the 111 strip shopping road segments between 2005 and 2009. Table 1 displays the number of crashes in each speed zone (columns) compared to the speed zone recorded in the Victorian crash data (rows). Almost one-quarter of the crashes (158/655) did not have the correct speed zone recorded.

Table 1. The number of crashes in each speed zone compared with the recorded speed zone

	True speed zone						
Speed zone recorded in crash data	**40 km/h**	**50 km/h**	**60 km/h**	**70 km/h**	**80 km/h**	**Total**	**Positive predictive value**
40 km/h	13	0	12	2	0	27	48.1%
50 km/h	2	7	51	0	2	62	11.3%
60 km/h	4	1	383	16	8	412	93.0%
70 km/h	0	0	10	41	15	66	62.1%
80 km/h	0	0	3	7	53	63	84.1%
Other	0	0	2	0	0	2	0%
Missing	2	1	15	4	1	23	0%
Total	21	9	476	70	79	655	
Sensitivity	61.9%	77.8%	80.5%	58.6%	67.1%		

Key: cells shaded in grey indicate crashes where speed zone was correctly recorded

Sensitivity of recorded speed zone

The sensitivity (that is, the probability that a crash that occurred in a particular speed zone was recorded as being that speed zone) of the speed zone recording varied by speed zone. Sensitivity was lowest for 70 km/h zones (58.6%) and highest for 50 km/h and 60 km/h zones (77.8% and 80.5%, respectively). Simple logistic regression (outcome=whether or not the speed zone was recorded correctly, predictor=true speed zone) revealed that crashes in 60 km/h zones were significantly more likely to have the speed zone recorded correctly than crashes in 40 (OR=2.53, 95% CI 1.02-6.29), 70 (OR=2.91, 95% CI 1.72-4.93) or 80 km/h zones (OR=2.02, 95% CI 1.20-3.40). There were no other statistically significant pairwise differences between speed zones in terms of the proportion of crashes in that speed zone for which the speed zone was correctly recorded.

Positive predictive value of recorded speed zone

PPV (the probability that a crash recorded as occurring in a speed zone actually occurred in that speed zone) was highest for crashes recorded as occurring in 60 km/h zones (93.0%) and 80 km/h zones (84.1%) but extremely low in 50 km/h zones (11.3%). Simple logistic regression (outcome=whether or not the speed zone was recorded correctly, predictor=recorded speed zone) revealed that PPV was significantly higher for crashes recorded as occurring in a 60 km/h zone than those recorded as occurring in all other speed zones (40 km/h, OR=14.22, 95% CI 6.12-33.08; 50 km/h, OR=103.77, 95% CI 43.37-248.29; 70 km/h, OR=8.05, 95% CI 4.31-15.04; 80 km/h OR=2.49, 95% CI 1.15-5.40). Likewise, PPV was significantly higher for crashes recorded as occurring in 80 km/h zones compared to crashes recorded as occurring all other speed zones except 60 km/h zones (40 km/h, OR=5.7, 95% CI 2.1-15.7; 50 km/h, OR=41.6, 95% CI 14.8-117.5; 70 km/h, OR=3.2, 95% CI 1.4-7.5). PPV was significantly lower for crashes recorded as occurring in a 50 km/h zone than those recorded as occurring in all other speed zones (40 km/h, OR=0.14, 95% CI 0.05-0.41; 60 km/h, OR=0.01, 95% CI 0.004-0.023; 70 km/h, OR=0.08, 95% CI 0.03-0.20; 80 km/h OR=0.02, 95% CI 0.01-0.07.).

Factors associated with accuracy of recorded speed zone

The speed zone was recorded correctly for 80.5% of the 441 crashes that police attended compared to 66.2% of the 213 crashes that they did not attend. The odds of the speed zone being recorded correctly more than doubled if the police attended the crash compared to when they did not (simple logistic regression, outcome=whether or not the speed zone was recorded correctly, predictor=whether or not police attended the crash scene; refer to Table 2, column 2).

The speed zone was recorded correctly for all of the nine crashes that had a fatal outcome. Of the 232 serious injury crashes (where the police reported that at least one road user was admitted to hospital), 78.9% had the speed zone recorded correctly, compared to 73.7% of the 305 other injury crashes (where the police reported that none of the injured road users were admitted to hospital). While there was some evidence that crashes in which someone was killed or seriously injured were more likely to have the speed limit recorded correctly than other injury crashes (simple logistic regression, outcome=whether or not the speed zone was recorded correctly, predictor=whether or not a road user was killed or seriously injured in the crash; refer to Table 2, column 2), this effect disappeared if both police attendance and accident severity were included in the same model (multiple logistic regression, outcome=whether or not the speed zone was recorded correctly, predictors=whether or not a road user was killed or seriously injured in the crash & whether or not police attended the crash scene, refer to Table 2, column 3). This is because police were more likely to attend crashes with a more severe outcome. Therefore, the attendance of police (which was associated with the severity of the crash) was associated with more accurate recording of speed zone.

Table 2. The association between police attendance, accident severity and correct recording of speed zone

	n (%) of crashes with speed zone recorded correctly	**Crude odds ratio (95% confidence interval)**	**Adjusted odds ratio (95% confidence interval)**
Police attended crash			
No	141 (66.2%)	1.0	1.0
Yes	355 (80.5%)	2.11 (1.46-3.05)	2.04 (1.38-3.02)
Road user killed or seriously injured			
No	305 (73.7%)	1.0	1.0
Yes	192 (79.7%)	1.40 (0.96-2.05)	1.10 (0.73-1.67)

Key: Column 2—Estimates from two separate simple logistic regressions (outcome=whether speed zone was recorded correctly, variables as defined in table rows); Column 3—Estimates from multiple logistic regression (outcome=whether speed zone was recorded correctly, both variables included in model)

Implications of speed zone misclassification: relationship between speed zone and crash severity

As a demonstration of the influence that misclassifying the speed zone of a crash can have on analyses that use these data, we investigated the association between speed zone and crash severity using simple logistic regression. The outcome of interest was whether any road user was killed or seriously injured in the crash. In the first model, we used the true speed zone to estimate the association between speed zone and crash severity. In the second model, we used the speed zone recorded in the crash data (of which almost 25% were misclassified) to estimate the association between speed zone and crash severity. Table 3 displays the results from each of the models. The odds ratios reflect the odds of a crash in that speed zone resulting in at least one road user who was killed or seriously injured, compared to the odds of a crash in 60 km/h zones resulting in one road user who was killed or seriously injured.

Table 3. The association between speed zone and accident severity: influence of misclassification of speed zone

	True speed zone		**Speed zone recorded in the crash data**	
Speed zone	**Odds ratio**	**95% CI**	**Odds ratio**	**95% CI**
40 km/h	**2.44**	**1.01-5.92**	1.62	0.74-3.54
50 km/h	0.92	0.23-3.71	0.66	0.36-1.10
60 km/h	1.0		1.0	
70 km/h	1.54	0.93-2.56	**1.86**	**1.10-3.13**
80 km/h	0.90	0.54-1.49	1.00	0.58-1.74

Key: $p<0.05$ indicated in bold type. Odds ratios=odds of a crash in that speed zone resulting in at least one road user who was killed or seriously injured, compared to the odds of a crash in 60 km/h zones resulting in one road user who was killed or seriously injured

If the odds ratios and 95% confidence intervals are compared across models for each speed zone, it is apparent that although there are some differences, the 95% confidence intervals for the estimates overlap (e.g. the 95% confidence intervals for 50 km/h are from 0.23 to 3.17 when the true speed zone is used, and from 0.36 to 1.10 when the speed zone recorded in the crash data is used). As such, it could be argued that the misclassification of speed zone does not have a large effect on the estimates of the association between speed zone and crash severity.

However, if the results of the two models are compared in terms of the speed zones that were significantly associated with a severe outcome, the true cost of misclassification of speed zones becomes apparent. When using the true speed zone, the odds of someone being killed or seriously injured in the crash were significantly higher in 40 km/h zones than in 60 km/h zones (OR=2.44, 95% CI 1.01-5.92) and 80 km/h zones (OR=2.72, 95% CI 1.02-7.27). There were no other statistically significant differences between speed zones when the true speed zone was used. In contrast, when the recorded speed zone was used, the odds of someone being killed or seriously injured were significantly higher in 70 km/h zones compared to 50 km/h zones (OR=2.81, 95% CI 1.34-5.88) and 60 km/h zones (OR=1.86, 95% CI 1.10-3.13). Therefore, the misclassification of speed zone in the police-reported crash data makes crashes in 70 km/h zones appear more severe than other speed zones, and hides the increased severity in 40 km/h zones in this sample of strip shopping centre arterial road segments in metropolitan Melbourne.

Conclusions

A comparison of the speed zone recorded for the location of a sample of urban arterial crashes in the Victorian crash database with the true speed zone revealed that the speed zone was not recorded correctly for almost one-quarter of the crashes occurring on midblock road segments.

The sensitivity of the speed zone data (that is, the proportion of crashes that occurred in a speed zone that were correctly recorded as occurring in that speed zone) was significantly higher for crashes that occurred in 60 km/h zones than all other speed zones. The majority of urban arterials in this study, and across Melbourne, have a speed limit of 60 km/h, which is likely to impact sensitivity.

PPV is an indicator of the trust that a data-user can place in the accuracy of the information recorded in the crash data. PPV also differed significantly across speed zones and was highest for crashes recorded as occurring in 60 km/h zones followed by crashes recorded as occurring in 80 km/h zones. The extremely low PPV for crashes recorded as occurring in 50 km/h zones is concerning. For example, almost 90% of crashes that were recorded as occurring in 50 km/h zones did not, in fact, occur in 50 km/h zones. The default speed limit for roads in built up areas in Victoria is 50 km/h. It is possible this is the reason that the attending police and/or the person who reported the crash at a police station erred in reporting that the crash occurred in a 50 km/h zone when in fact it did not. This highlights the inherent potential for recall bias when people are asked to recall the circumstances of the crash.

Another finding from this study that indicates that recall bias is an issue is that the speed zone of the crash was significantly more likely to be recorded correctly if the police attended the crash than if they did not. Police experienced at completing crash report forms will be aware that they need to identify the speed zone at the crash location and can investigate to determine what speed zone the crash is in if/when they attend the scene. They are therefore less likely to need to rely on their memory or an educated guess regarding the speed zone, unlike people that report a crash at a police station, who may not be aware what information they need to provide.

Finally, a demonstration analysis showed that the misclassification of speed zone in this sample of crashes from the Victorian crash data influenced the results of a simple uncontrolled study into the relationship between speed zone and crash severity. The use of the misclassified speed zone data led to estimates of association that identified crashes recorded as being in 70 km/h zones as significantly more likely to lead to a severe crash outcome than those crashes recorded as being in 50 km/h or 60 km/h zones. In contrast, the use of the true speed zone data indicated that the crash outcome was significantly more likely to be severe in 40 km/h zones than in 60 km/h zones or 80 km/h zones. It must be emphasized that this analysis on a relatively small sample of road segments did not adjust for any potential confounders and hence the results may be explained by other factors that differed between road segments apart from speed zone. It does, however, demonstrate that using data that are misclassified can lead to different conclusions compared to when accurate data are used. These different conclusions would lead to different recommendations for reducing crashes and improving road safety which underscores the importance of ensuring accurate data collection in sources that are used to inform real-world policy and practice.

There are several limitations to this study. First, it is not known how accurate the true speed zone data provided by Vicroads is, although Vicroads staff did refer to historical records to establish that the speed zone did not change over the study period on these roads. The sample was restricted to a relatively small number of complex urban roads that were strip shopping zones; it is unknown if the accuracy of the recording of speed zones would differ in other urban areas or rural areas. Finally, the study excluded crashes occurring at intersections because there was no information on the speed zone of the intersecting road, however, this means the results cannot be generalised to crashes that occur at intersections. This study, however, can be considered a pilot study and provide impetus for future research to establish the accuracy of the recording of speed zone in the police-reported crash data for all roads in Victoria.

We have demonstrated that the recording of speed zone in the Victorian crash data can be inaccurate and that the sensitivity and positive predictive value varies by speed zone. We have also demonstrated that this can influence the results of investigations that use the misclassified data, e.g. when identifying risk factors for crash severity. The methods for collecting the crash data are subject to human error and recall bias. There is, however, potential to improve the accuracy of the data that are included in the Victorian crash data by taking advantage of the opportunity to spatially link the crash data (using the location of the crash) to other existing data sources. The Victorian Government data directory (data.vic.gov.au) has a vast amount of data that could be used for this purpose, for example, the speed zone maps for Victorian roads. Other geo-spatially coded data of road infrastructure and assets could also be used to collect accurate data about the crash location. This approach of linking crash data to speed zone maps is already in use in Western Australia. This approach would both improve the accuracy of the data, and lessen the administrative burden on police officers of collecting a large amount of data that could be collected by other means.

References

Boufous, S., de Rome, L., Senserrick, T. & Ivers, R. (2012). Risk factors for severe injury in cyclists involved in traffic crashes in Victoria, Australia. *Accident Analysis and Prevention, 49*, 404-409

Buckis, S., Lenne, M.G., Stephens, A., Bingham, C.R. & Fitzharris, M. (2016). Young driver crash types and lifetime care costs by posted speed limit. *Injury Prevention, 22* (suppl 2), A121

Garratt, M., Johnson, M. & Cubis, J. (2015). Road crashes involving bike riders in Victoria, 2002-2012: an Amy Gillett Foundation report. Downloaded July 16, 2017 from http://www.amygillett.org.au/wp-content/uploads/2015/09/Road-crashes-AGF-Report-FINAL-Sept-2015.pdf

Hennekens, C.H. & Buring, J.E. (2010). *Epidemiology in Medicine* (1st ed.). Boston, USA: Little, Brown & Co.

Newstead, S.V., Watson, L.M. & Cameron, M.H. (2013). *Vehicle Safety Ratings Estimated from Police Reported Crash Data: 2013 Update Australian and New Zealand Crashes During 1987-2011*. Monash University Accident Research Centre, Report No. 318.

Stephan, K.L. (2015). *A Multidisciplinary Investigation of the Influence of the Built Urban Environment on Driver Behaviour and Traffic Crash Risk.* Monash University ethesis, available from https://doi.org/10.4225/03/58b78d6fea3f1

VicRoads (2008). *CrashStats User Guide Road Crash Statistics*. Kew, Victoria, Australia: VicRoads

Vicroads (2016). 2015 Victorian Road Trauma: Analysis of Fatalities and Serious Injuries. Downloaded July 16, 2017 from https://www.vicroads.vic.gov.au/safety-and-road-rules/safety-statistics/crash-statistics

World Health Organization (2010. Data Systems: A Road Safety Manual for Decision-makers and Practitioners. Geneva Switzerland: World Health Organization.